BEYOND REWARD MAXIMIZATION: EVALUATING THE DIVERSITY OF TRAJECTORIES IN REINFORCEMENT LEARNING WITH TEMPORAL VENDI SCORE

Tom Stanic[*1], **Marco Jiralerspong**[1,2], **Xiaofeng Zhang**[1,2], **Danilo Vucetic**[1,2], **Gauthier Gidel**[1,2,3]

[1]Université de Montréal [2]Mila [3]CIFAR

ABSTRACT

In domains such as scientific discovery and automated design using reinforcement learning (RL), the final task of an agent should extend beyond maximising a single scalar reward; it requires identifying diverse sets of high-quality trajectories to uncover distinct solutions that can provide novel insights on how to solve the problems of interest and transfer robustly from simulation to the real world. However, the RL literature currently lacks a holistic, domain-agnostic standard for measuring trajectory diversity. Existing metrics have been developed to improve exploration at training time but not to evaluate and compare diversity induced by different agents, rendering cross-method comparisons inconsistent and challenging. To address this, we introduce the Temporal Vendi Score (`TVS`), a novel metric designed to evaluate the diversity of an RL agent by computing the entropy of the eigenvalues' similarity matrix of sampled trajectories. Unlike previous approaches, our metric captures the behavioral diversity of trajectories by accounting for both the sequential nature of state visitations and the temporal structure of the underlying MDP, rather than relying on order-agnostic state comparisons. We validate the `TVS` on simple environments where we can control the number of different ways a problem can be solved, demonstrating that it provides a more robust, semantically meaningful ranking of diversity than standard baselines. We then show that our metric can scale to a high-dimensional, continuous environment.

1 INTRODUCTION

Recent work in RL has highlighted that the goal is often not merely to maximize a single scalar objective, but also for the agent to identify *diverse* high-quality solutions and skills. A growing body of work has therefore incorporated diversity objectives directly into training, e.g., maximum-entropy (Nachum et al., 2017; Haarnoja et al., 2017; 2018; Lee et al., 2019), unsupervised skill discovery (Eysenbach et al., 2019; Sharma et al., 2019; Campos et al., 2020), exploration (Pathak et al., 2017; Burda et al., 2018; Ecoffet et al., 2019; Guo et al., 2022), quality-diversity trade-off (Mouret & Clune, 2015; Pugh et al., 2016; Lim et al., 2023), etc. Although implemented with different algorithms and loss functions, these methods share a common motivation: learning a diverse set of behaviors can improve downstream performance and usability, including generalization, robustness, exploration efficiency, etc. Similar needs for multiple distinct solutions also arise in adjacent areas such as large language model reasoning (Yao et al., 2025; Yu et al., 2025), diverse content generation (Zhang et al., 2026; Trang et al., 2025), and scientific discovery problems (Jain et al., 2022; Zhu et al., 2023), where different trajectories can correspond to different valid solutions or candidates.

Despite the growing interest in diversity in RL, the field lacks a comprehensive and general metric to quantify the diversity of learned policies. The focus on reward maximization has trickled down to the evaluation metrics used in RL: the dominant evaluation protocol remains average return. Even in papers explicitly designed to encourage diversity, existing proxies fall short because they often

[*]Correspondence to `tom.stanic@mila.quebec`

measure something adjacent to diversity rather than diversity itself. For instance, policy entropy can be increased by injecting action noise without producing meaningfully different outcomes; state visitation counts depend heavily on the choice of representation and discretization; and goal-coverage metrics require environment-specific structure that may not exist in general continuous-control tasks. As a result, it is often unclear *how diverse* a learned policy actually is, and whether improvements in return are driven by truly broader behavioral repertoires or by unrelated changes in optimization and exploration.

In this work, we formalize diversity at the level of *agent behavior* in terms of how many different trajectories that a trained policy can sample, and propose an interpretable metric, the temporal Vendi score (`TVS`). Concretely, the diversity of a trained policy is evaluated by how varied the induced distribution of trajectories is. Given two sampled trajectories, we propose a time-to-reach distance to quantify the cost of transitioning between states selected respectively in each trajectory. Then, we use the goal alignment kernel (GAK) (Cuturi et al., 2007) that marginalizes over the state pairs to get the similarity between the trajectories. Finally, we leverage the Vendi score (Friedman & Dieng, 2023; Pasarkar & Dieng, 2024) to aggregate the similarity matrix among all sampled trajectories into a scalar interpretable value of diversity.

We evaluate our proposed metric on maze tasks spanning discrete and continuous settings where we know the ground truth possible trajectories. Our experimental results show that `TVS` has several desirable properties, including 1) interpretability as the raw value aligns with human perceptions, 2) sample efficiency as the metric converges with reasonable amounts of sampled trajectories, 3) scalability as the computation is parallelizable using GPUs, 4) and universality as it can be applied to both discrete and continuous setups. We summarize our contributions as follows.

Contributions:

1. We propose a novel, interpretable diversity metric for RL agents that is sensitive to both the order of selected actions and the geometry of the underlying MDP.
2. We show the metric is able to identify the diversity of agent behavior in toy tasks, agrees with human intuition and correlates better with increases in diversity than coverage/entropy-based measures.
3. We demonstrate that the metric converges with relatively few sampled trajectories and that it has the potential to scale to larger environments with continuous state/action spaces.

2 Background

2.1 Reinforcement learning

The focus of reinforcement learning on maximizing expected return is almost foundational. In fact, Sutton & Barto (2018) defines RL as learning a policy (i.e. a mapping from states to actions) "so as to maximize a numerical reward signal". More formally, RL considers Markov decision processes (MDPs) as the main object of study. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P}$ is the state-transition probability function, $\mathcal{R}$ is the reward function and γ is the discount factor. The goal is then to learn a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes the expected sum of discounted rewards given by $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$.

2.2 Vendi score

While evaluation of RL has focused on average return, generative model evaluation has already progressed to diversity metrics. Among them, Friedman & Dieng (2023) propose using the Vendi score to evaluate diversity. Inspired by ecological diversity metrics, the Vendi score is a sample-based metric that takes a similarity function k. Then, for a set of n samples $x_1, \ldots, x_n$ the similarity function must have $k(x, x) = 1$ for all samples. The Vendi score computes $\mathbf{K}$, the kernel matrix of pairwise similarities. Finally, the Vendi score of the set of samples is the entropy of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix $\mathbf{K}/n$: $\mathrm{VS}(x_1, \ldots, x_n) = \exp - \sum_{i=1}^{N} \lambda_i \log(\lambda_i)$ (with $0 \log 0 = 0$ for $\lambda_i = 0$). The Vendi score has many desirable properties. In particular, it can be roughly interpreted as a continuous estimator of the number of completely dissimilar elements.

Table 1: Comparison of properties of existing diversity metrics. †Applicable to a single agent's trajectories, without requiring a multi-agent population or skill-conditioned policy. Note that DIAYN requires a latent skill-conditioned policy and thus cannot be applied to an arbitrary single agent.

Metric	Continuous States	Sequentiality	Single Agent†	Population	Temporal MDP Structure
Policy entropy	×	×	✓	×	×
State entropy	×	×	✓	✓	×
DIAYN (Eysenbach et al., 2019)	✓	✓	×	✓	∼
Pairwise DTW (Müller, 2007)	✓	✓	✓	×	✓
SND (Bettini et al., 2025)	✓	×	×	✓	×
DvD (Parker-Holder et al., 2020)	✓	×	×	✓	×
TVS (ours)	✓	✓	✓	✓	✓

2.3 Quasimetric functions

Crucial to the computation of the Vendi score is the choice of similarity function. To have such a function between RL trajectories, we can turn to computing a notion of distance between states that leverages temporality (i.e. how many timesteps would it take to get from state s to state s'). Quasimetric functions (a generalization of a metric function that does not require symmetry) are a natural choice for such functions. A quasimetric is a function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ satisfying non-negativity, identity ($d(s, s) = 0$), and the triangle inequality ($d(s_1, s_2) + d(s_2, s_3) \geq d(s_1, s_3)$), but crucially not requiring symmetry.

In goal-conditioned RL, where we are interested in getting to some goal state as quickly as possible, the optimal goal-conditioned value function $V^*(s|s')$ is necessarily a quasimetric (Wang et al., 2023). If the cost we ascribe to performing a transition is 1, this optimal value function exactly gives us the number of timesteps it would take to go from s to s'. This quasimetric function can be learned in a tabular manner (e.g. by value iteration) or with a function approximator.

In particular, Wang et al. (2023) use a model $d_\theta(s, s')$ that is trained by maximizing $\mathbb{E}_{s\sim p_{\text{state}}, g\sim p_{\text{goal}}}[d_\theta(s, g)]$ subject to the constraint that local costs are not overestimated: $d_\theta(s, s') \leq -r$ for all observed transitions (s, a, s', r) where a represents an action and r represents the reward. This objective pulls states apart globally while respecting observed transition costs locally, effectively learning the geometry induced by the environment's dynamics. With a reward of -1 for each transition, d_θ learns an approximation of the time-to-reach distance, i.e. the minimum number of steps to go from s to g. More recent work has also explored learning temporal distances via contrastive successor features (Myers et al., 2025), which could further improve the accuracy and stability of such estimates.

3 Related Work

Diversity is frequently used as a training signal rather than a formal evaluation metric. Commonly, this training signal is given by a loss that balances maximizing entropy (either at the policy (Haarnoja et al., 2017; Nachum et al., 2017) or state level (Hazan et al., 2019; Ashlag et al., 2025)) and the return of the agent. While diversity can be quantified by examining the average entropy of the policy over visited states (Haarnoja et al., 2018), this often measures stochasticity (randomness) rather than distinct behavioral modes or strategies. Similarly, Hazan et al. (2019) use state-visitation entropy to proxy state-space coverage. However, this requires discretization and discards temporal sequential ordering. To capture sequential structure, trajectory similarity measures, e.g., Dynamic Time Warping (DTW), have been used to measure trajectory similarity in imitation learning contexts (Vakanski et al., 2012; Wu et al., 2020). While effective for alignment, these measures focus on similarity to a template rather than inherent diversity within a policy.

Beyond entropy-based measures, learned metrics have also been used to encourage diversity in skill discovery methods. For example, Diversity is All You Need (DIAYN) (Eysenbach et al., 2019) train a discriminator network to classify which skill latent an agent used to generate a given trajectory. The classification accuracy can be repurposed to quantify diversity in behavioral signatures as higher

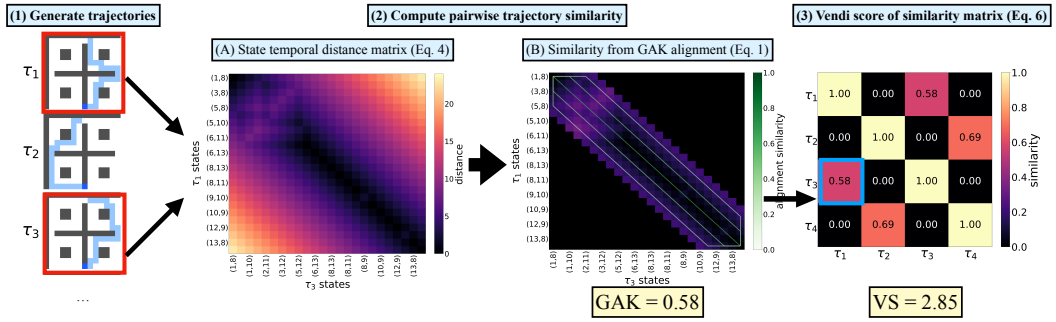

Figure 1: Procedure for computing the temporal Vendi score (TVS). (1) Trajectories are sampled from the agent. (2A) For each pair of trajectories, a pairwise state temporal distance matrix is computed whose entries correspond to the symmetrized time-to reach distance between a pair of states s_i, s'_j, see Eq. 3. (2B) A truncated band of this matrix is used to compute the GAK alignment using Eq. 1. The resulting GAK is used as similarity value between that pair of trajectories. (3) TVS is computed from the eigenvalues of the resulting pairwise similarity matrix using Eq. 6.

accuracy indicates a distribution that is easier to distinguish. However, as this metric relies on a neural network classifier that needs to be trained separately for each agent and requires an agent to have a latent skill-conditioned policy, it cannot be applied universally as a metric.

In multi-agent settings, metrics like System Neural Diversity (SND) (Bettini et al., 2025) and Role Diversity (RD) (Hu et al., 2022) aggregate pairwise distances between policies. However, pairwise approaches are susceptible to clustering—where populations split into redundant groups. To address this, Parker-Holder et al. (2020) introduced Diversity via Determinants (DvD), which uses a kernel matrix determinant to measure the volume spanned by the population. While DvD captures global redundancy, its reliance on embeddings from randomly sampled states fails to account for the temporal dependencies or sequential structures critical in spatially-structured tasks. In general, methods applicable to multi-agent settings/populations focus on evaluating diversity between policies but cannot evaluate behavioral diversity within a single policy.

Alternatively to determinants and matrix ranks, one can consider the Vendi score (Friedman & Dieng, 2023) and its variants (Pasarkar & Dieng, 2024) as aggregation method. While the metric has been adopted for use in evaluation of generative models, Lintunen (2025) use the Vendi score to measure diversity and use that as reward to train agents with diverse skills. However, their similarity kernel used to compute the Vendi score relies on simple trajectory-based statistics (such that as the mean and covariance matrix of trajectory observations) and it does not take into account the temporal geometry of the underlying environment. The latter issue arises from their metric only considering the euclidean distance between states. Problematically, two states can be close (coordinate-wise) while being far apart in practical terms for an agent (for example if there is a long wall between them). By considering the time-to-reach distance between states, our metric is aware of the underlying dynamics of the MDP.

4 METHOD

Inspired by how species diversity is computed in ecology and how diversity is evaluated in generative modeling, we adopted the widely used Vendi score (Friedman & Dieng, 2023) that computes diversity by using similarity values between instances. Our core contribution is to evaluate the similarity of RL trajectories by using a time-to-reach distance that captures the temporal state distances between states visited by the trajectories while considering the alignment of these trajectories. As illustrated in Fig. 1, our approach comprises three key steps. First, we sample trajectories from an agent or a population of agents through environment rollouts. Second, we compute the cost function, i.e., the proposed distance, between all states given two sampled trajectories, which reflects the geometry of the environment (§4.1.3). Then, we adopt the global alignment kernel (GAK) (Cuturi et al., 2007) to get the pairwise trajectory similarities (§4.1). Finally, we aggregate this similarity matrix between trajectories into a single diversity value using the Vendi score.

4.1 TRAJECTORY SIMILARITY

To apply the vendi score in RL contexts first requires selecting the right objects to apply it to. We propose comparing **trajectories**. In doing so, we address many of the desiderata mentioned in Tab. 1. We can evaluate the diversity of a single agent or a population by sampling a set of rollouts and evaluating the diversity of those trajectories. In addition, comparing trajectories ensures that sequentiality is taken into account: agents that visit similar states but in different orders are considered distinct. Finally, by using the time-to-reach distance, we can ensure the metric reflects the transition dynamics of the underlying environment. Below we detail our method for evaluating the similarity of trajectories.

4.1.1 GLOBAL ALIGNMENT KERNEL

As trajectories can be interpreted as sequences of states, we take advantage of the rich literature on distances between sequences and adopt the Global Alignment Kernel (GAK) (Cuturi et al., 2007), which compares two sequences by marginalizing over all valid monotonic alignments

$$k_{\mathrm{GA}}(\tau, \tau') = \sum_{\pi \in \mathcal{A}(T,T')} \prod_{(i,j)\in\pi} \kappa(s_i, s'_j), \tag{1}$$

where $\mathcal{A}(T, T')$ is the set of all valid monotonic alignments between sequences of lengths T and T', and $\kappa(s_i, s'_j) = \exp(-d(s_i, s'_j)/\sigma)$ is a local similarity kernel between sequence elements, with d a distance and $\sigma > 0$ a bandwidth parameter. GAK is guaranteed to be positive semi-definite (PSD) whenever $\kappa/(1+\kappa)$ is itself PSD (Cuturi et al., 2007) and can be computed efficiently via the recursion

$$\mathrm{GA}(i, j) = \kappa(s_i, s'_j) \cdot (\mathrm{GA}(i-1, j-1) + \mathrm{GA}(i-1, j) + \mathrm{GA}(i, j-1)) \,. \tag{2}$$

In practice, we restrict alignments to a Sakoe–Chiba band (Sakoe, 1978) which limits i, j to a band around the diagonal given by

$$|i - j| \leq 0.2 \cdot \max(T, T'),$$

reducing the complexity to $\mathcal{O}(T \cdot 0.4T')$. Intuitively, trajectories that differ by more than 20% in temporal alignment are already behaviorally distinct, and will contribute to a high `TVS` regardless of the exact alignment. This assumption is reasonable in practice: trajectories collected from a trained agent tend to share broadly similar lengths, as the agent has learned to act purposefully rather than wandering arbitrarily. Any misalignment exceeding the band width therefore reflects genuine behavioral divergence rather than noise, so the band discards only alignments that would correspond to already-diverse trajectory pairs, without losing discriminative information for similar ones. We further verify this in Appendix A.3.

4.1.2 ENVIRONMENT GEOMETRY VIA TIME-TO-REACH DISTANCE

The local cost $d(s, s')$ used by the kernel κ in GAK should reflect meaningful behavioral differences grounded in the environment's dynamics, rather than arbitrary distance in observation space while also being broadly applicable to different RL problems. We propose to set the cost to the *time-to-reach* distance: the minimum number of actions required to reach state s' from state s (Steccanella & Jonsson, 2022; Wang et al., 2023; Park et al., 2023). Two states are close if an agent can quickly transition between them, regardless of their distance in the observation space. We note that while this distance does not guarantee that GAK is PSD, we observe that eigenvalue violations are negligible (on the order of 10^{-5}) and clamp them to zero. We verify in Section A.3 that the metric remains stable under perturbation of these small negative eigenvalues.

Exact computation in discrete environments. In grid-world environments with discrete state spaces, time-to-reach distances can be computed exactly via breadth-first search (BFS). We identify walkable cells by excluding walls, closed doors, and other non-traversable obstacles from the environment grid. For each walkable cell, BFS explores the neighborhood (up, down, left, right) to compute shortest paths to all other reachable cells, with each move incurring unit cost. This yields a distance matrix $D \in \mathbb{R}^{N \times N}$ where D_{ij} is the minimum number of steps between cells i and j.

Learned approximation via quasimetric models. When the state space becomes large or continuous, exact computation becomes infeasible. We employ learned quasimetric models (Wang et al., 2023), which learn the optimal goal-conditioned value function $V^*(s; g)$ for any Markov decision process (MDP). Since quasimetric distances are asymmetric in general (reflecting directional dynamics such as descending versus ascending a hill), we symmetrize the distance d by averaging both directions, yielding a single undirected cost between any two states:

$$d_{\text{sym}}(s, s') = \frac{1}{2}\big(d(s, s') + d(s', s)\big). \tag{3}$$

This yields a symmetric local kernel

$$\kappa(s_i, s'_j) = \exp\left(-\frac{d_{\text{sym}}(s_i, s'_j)}{\sigma}\right). \tag{4}$$

Details on the calibration of σ from random rollout statistics are provided in Appendix A.1. Once trained, the symmetrized distances can be used to pre-compute, for each pair of trajectories, the pairwise state cost matrix required by the GAK recursion (Eq. 2). Since trajectory comparisons are independent, all N^2 dynamic programming problems can be parallelized on GPU.

4.1.3 Normalization of similarity matrix

By applying the GAK and time-to-reach distance, we get the similarity matrix between all sampled trajectories. However, the raw GAK similarity $k_{\text{GA}}(\tau_i, \tau_j)$ does not satisfy $k_{\text{GA}}(\tau, \tau) = 1$, which is required by the Vendi score. We therefore normalize the kernel matrix using

$$K_{ij} = \frac{k_{\text{GA}}(\tau_i, \tau_j)}{\sqrt{k_{\text{GA}}(\tau_i, \tau_i) \cdot k_{\text{GA}}(\tau_j, \tau_j)}}, \tag{5}$$

yielding $K_{ii} = 1$ for all i.

4.2 Temporal Vendi Score

Given a set of N trajectories, the procedure described above yields a normalized similarity matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ with $K_{ij} \in [0, 1]$ and $K_{ii} = 1$. We aggregate this matrix into a single scalar diversity measure using the Vendi score (Friedman & Dieng, 2023). In its original form ($q = 1$), each eigenvalue is weighed proportionally to its magnitude. Pasarkar & Dieng (2024) generalize this to a family of orders q.

Higher orders ($q > 1$) concentrate on the dominant eigenvalues, effectively counting only the major behavioral modes while filtering out small variations that do not represent genuinely different strategies. We use $q = 2$ throughout our experiments. This choice focuses the score on the dominant eigenvalues of $\mathbf{K}/N$, effectively counting major behavioral modes while discounting superficial diversity from action-level stochasticity. It also converges with fewer trajectory samples (Pasarkar & Dieng, 2024) and is more robust to the negligible negative eigenvalues discussed in Section 4.1.

The temporal Vendi score (`TVS`) is given by

$$\texttt{TVS}(\{\tau_1, \ldots \tau_N\}) = \exp\left(-\log \sum_{i=1}^{k} \lambda_i^2\right), \tag{6}$$

where $\{\lambda_1, \ldots, \lambda_k\}$ are the eigenvalues of the normalized similarity matrix $\mathbf{K}/N$.

5 Experiments

The experimental validation of our metric takes advantage of simple mazes where we know the ground truth diverse trajectories that can be taken to reach the goal/goals. We aim to demonstrate the following: (1) our metric provides an interpretable evaluation of the diversity of an agent, (2) it converges with a tractable number of samples, (3) it is more sensitive to genuine increases in diversity than baseline metrics and (4) it can scale to larger continuous environments.

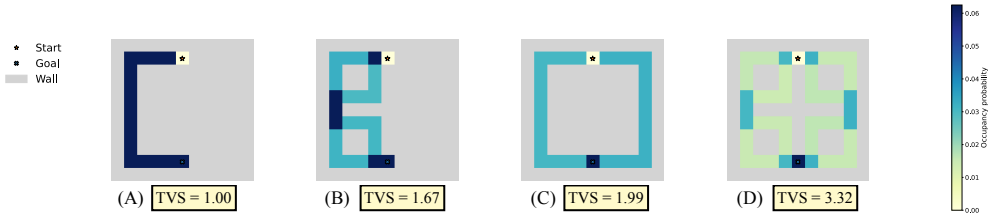

Figure 2: TVS of agents with increasing diversity in a simple grid maze. The maze admits paths on both its left and right sides, each with minor variants. (A) The agent follows a single path to the goal. (B) The agent takes four overlapping paths, all on the left side of the maze. (C) The agent takes two paths, one on each side of the maze. (D) The agent takes eight overlapping paths, four on each side, mirroring the configuration of (B) on both sides.

5.1 Fundamental Properties

We use a toy maze set up with verifiable ground truth diversity to verify the fundamental properties of our proposed metric. To generate rollouts with controlled diversity, we compute a soft value iteration policy (Sutton & Barto, 2018; Geist et al., 2019) with a temperature parameter $\alpha = 0.002$, which induces a near-optimal stochastic policy that distributes probability mass across multiple solutions. By selectively filtering which paths the agent is allowed to take, we construct four settings with increasing diversity.

Interpretability. We begin by verifying the interpretability of our metric. As a Vendi score of n reflects a diversity equivalent to having n completely distinct objects, we aim for TVS to roughly reflect the number of distinct trajectories, accounting for their temporal separation in the environment. As shown in Fig. 2, TVS returns exactly 1 for a single-path agent (A), confirming no diversity. When four overlapping paths all stay on the same side (B), the score is only 1.67, correctly discounting their geometric proximity. Two paths on opposite sides (C) yield 1.99, nearly matching the ground-truth count of 2. Finally, mirroring (B) onto both sides (D) gives 3.32, closely matching $2 \times 1.67 = 3.34$ and confirming that the metric scales consistently with duplicated structure.

Sample efficiency. We now verify that TVS converges to a stable score with a moderate number of rollouts. Fig. 3 shows the convergence of TVS across four environments of varying size and structure: three single-goal environments (`single-goal small`, `single-goal medium`, `single-goal big`) and one multi-goal environment `multi-goal`. In all cases, the score stabilizes within 128 rollouts, demonstrating practical sample efficiency. The converged scores are interpretable: in the single-goal environments, the agent has access to two main paths, each with four minor variants, yielding a score of approximately 3 — reflecting two distinct strategies plus partial credit for the similar variants. In the multi-goal environment, eight distinct paths are available, and TVS converges to 7.6, closely tracking the number of genuinely different strategies. For additional environment configurations, see Appendix A.2.

5.2 Comparison with Baselines

To isolate the value of temporal structure in diversity measurement, we design an environment where state-level metrics are blind to meaningful behavioral differences. The environment requires the agent to pick up one of five keys and open a corresponding door to reach the goal.

We train agents with increasing diversity represented by the number of keys it learns to use (from 1 to 5) and measure the relative increase in each metric compared to the single-key baseline. As shown in Fig. 4, both state coverage and entropy of the occupancy measure remain nearly flat as the agent learns to solve the task with more keys. Coverage increases by less than 1% from 1 to 5 keys, and occupancy entropy similarly rises only marginally. This poor sensitivity demonstrates that state-level aggregation discards precisely the information needed to distinguish meaningfully different behaviors.

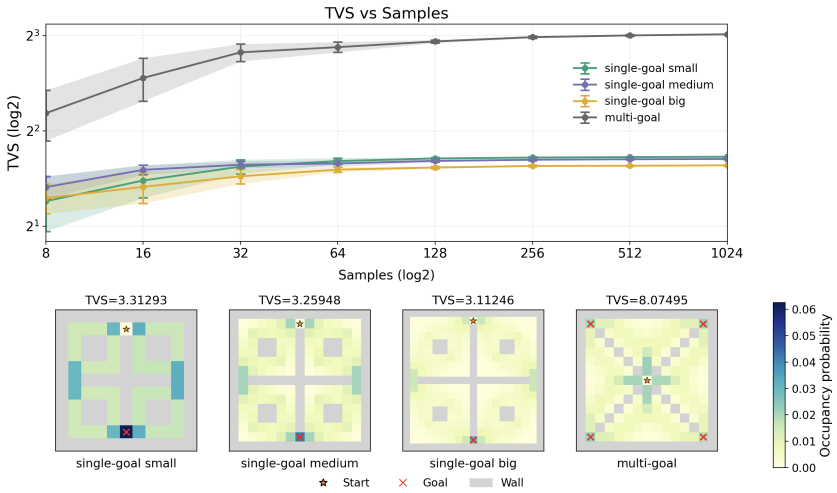

Figure 3: Convergence of TVS (with Sakoe Chiba 20%) as a function of the number of sampled trajectories across four environments of varying size and structure. State visitation heatmaps depict the full set of 1024 rollouts used for evaluation. TVS stabilizes within approximately 128 rollouts in all environments and converges to interpretable values reflecting the number of distinct strategies available.

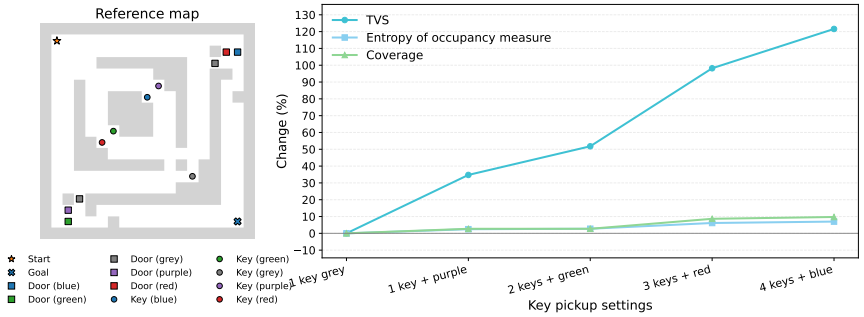

Figure 4: Relative increase in diversity score as the number of keys used grows from 1 to 5. State coverage and occupancy entropy remain nearly flat because the added diversity is temporal, the agent visits similar states in different orders. TVS captures this sequential structure and increases steadily with the number of distinct strategies.

These metrics fail because the added diversity is temporal, not spatial: the agent visits the same cells but in different orders. In contrast, TVS captures this sequential structure through trajectory alignment and its value increases substantially as the agents learn to use more keys, correctly reflecting that the agent has learned qualitatively distinct strategies.

5.3 Continuous Control

To demonstrate the potential scalability of our metric, we test it in the Maze2D environment introduced by Fu et al. (2020). The environment consists of a physics simulator where a ball with 2 degrees of freedom is force-actuated in the cartesian x/y directions. Agents must learn to control this ball and traverse a maze with a given structure and reach a goal. We train PPO (Schulman et al., 2017) agents with increasing entropy coefficients in mazes that replicate the structure of the discrete mazes (`single-goal`). The agents are trained with a dense reward corresponding to the negative

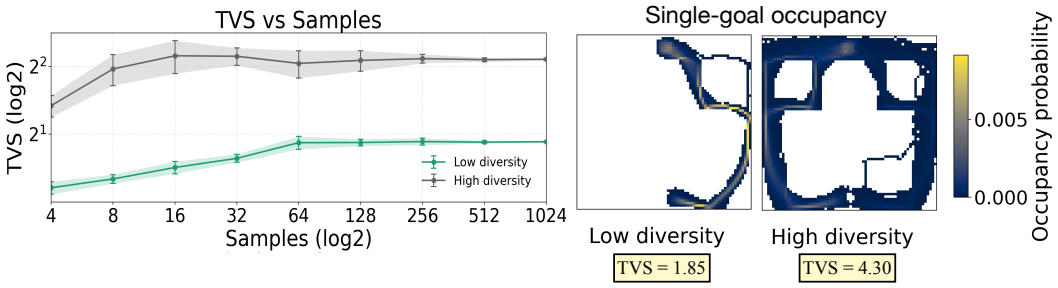

Figure 5: Convergence of TVS in continuous point maze environments using learned quasimetric distances, averaged over 4 seeds. We evaluate a low-diversity and a high-diversity agent trained via PPO with different entropy coefficients.

euclidean distance to the goal. We give large reward for actually reaching the goal to prevent the agent from learning to idle near the goal.

Since we know the general possible ground truth trajectories in these mazes, we visually inspect trajectories to match the trained agents to diversity buckets. For the low diversity agent, we pick an agent that almost always takes the same path. For the high diversity agent, we combine the trajectories of two agents that visually learn multiple trajectories to the goal for a given side. This need to combine trajectories comes from PPO agents heavily favoring one side or the other, even at higher entropy values.

To compute the TVS, we train a quasimetric model using the code provided by Wang et al. (2023), adapting it to the point maze environment. We train the quasimetric model offline, using the trajectories generated by the PPO agents. We filter out trajectories exceeding 300 timesteps, as these correspond to agents that fail to reach the goal efficiently and would inflate computation without contributing meaningful behavioral diversity.

As illustrated in Fig. 5, TVS converges even in continuous settings. The metric stabilizes at approximately 256 rollouts and produces scores consistent with the number of visually distinct strategies (with the low diversity agent learning two overlapping trajectories and the high diversity agent learning almost all possible paths to the goal). While these experiments are preliminary, they demonstrate the potential applicability of TVS to general RL domains.

Computational optimizations. In continuous state spaces, each state pair distance is obtained from a forward pass from the quasimetric model. For N trajectories of mean length T with a 20% Sakoe–Chiba band, this amounts to $\mathcal{O}(N^2 \cdot T \cdot 0.4T)$ distance evaluations. Even with batched GPU inference (approximately 5×10^5 pairs per second), this remains costly for large N and T.

To remedy this issue, we exploit the low-dimensional structure of the (x, y) point-maze observations by rounding states to a grid of resolution 0.1 and precomputing all pairwise distances on this grid once. Each subsequent GAK evaluation then requires only table lookups (see Fig. 10). We verify in Appendix A.4 that this discretization does not affect the scores. For higher-dimensional observations such as images, the number of unique states grows exponentially with dimension, making grid discretization impractical. Practical alternatives include clustering states in a learned embedding space to construct a finite lookup table, subsampling trajectories, or tightening the Sakoe–Chiba band (see Appendix A.4).

6 Conclusion

We introduced the temporal Vendi score (TVS), a domain-agnostic metric for evaluating the behavioral diversity of reinforcement learning agents. By combining time-to-reach distances that capture the geometry of the underlying MDP, the global alignment kernel that reflects the sequential structure of trajectories, and the Vendi score that aggregates pairwise similarities into an interpretable scalar, TVS addresses key shortcomings of existing diversity proxies. Our experiments on discrete maze environments with known ground-truth solutions show that TVS produces scores that align

with human intuition about the number of distinct strategies, converges within 128–256 sampled rollouts, and captures temporal diversity that state-level metrics such as coverage and occupancy entropy fail to detect. We further demonstrated that the metric scales to continuous control settings by leveraging learned quasimetric models and GPU-parallelizable computation.

Several directions remain for future work. While we take a first step towards scaling the metric, high-dimensional observation spaces such as pixel-based environments poses challenges for pre-computing pairwise distances. Furthermore, while we focus on distances between states, it could be interesting to leverage the actions taken in a trajectory as well as the reward obtained by the agent to further diversity evaluation. This extra information could be key for example in evaluating distinct locomotion gaits in robotics. Finally, an exciting direction is to use `TVS` not only as an evaluation metric but as a training signal to directly optimize for behavioral diversity, potentially complementing existing entropy-based and skill-discovery objectives.

REFERENCES

Yonatan Ashlag, Uri Koren, Mirco Mutti, Esther Derman, Pierre-Luc Bacon, and Shie Mannor. State entropy regularization for robust reinforcement learning. *arXiv preprint arXiv:2506.07085*, 2025.

Matteo Bettini, Ajay Shankar, and Amanda Prorok. System neural diversity: Measuring behavioral heterogeneity in multi-agent learning. *Journal of Machine Learning Research*, 26(163):1–27, 2025.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International conference on machine learning*, pp. 1317–1327. PMLR, 2020.

Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pp. II–413. IEEE, 2007.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes, 2019. URL https://arxiv.org/abs/1901.11275.

Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35: 31855–31870, 2022.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR, 09–15 Jun 2019.

Siyi Hu, Chuanlong Xie, Xiaodan Liang, and Xiaojun Chang. Policy diagnosis via measuring role diversity in cooperative multi-agent rl. In *International Conference on Machine Learning*, pp. 9041–9071. PMLR, 2022.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with GFlowNets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato

(eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9786–9801. PMLR, 17–23 Jul 2022.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, and Songhwai Oh. Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv preprint arXiv:1902.00137*, 2019.

Bryan Lim, Manon Flageat, and Antoine Cully. Understanding the synergies between quality-diversity and deep reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1212–1220, 2023.

Erik M. Lintunen. VendiRL: A framework for self-supervised reinforcement learning of diversely diverse skills. In *Workshop on Scaling Environments for Agents*, 2025. URL https://openreview.net/forum?id=cwwzYlTsfw.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making, 2025. URL https://arxiv.org/abs/2406.17098.

Meinard Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, 2:69–84, 01 2007. doi: 10.1007/978-3-540-74048-3_4.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned rl with latent states as actions. In *ICML workshop on new frontiers in learning, control, and dynamical systems*, 2023.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof M Choromanski, and Stephen J Roberts. Effective diversity in population based reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18050–18062. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d1dc3a8270a6f9394f88847d7f0050cf-Paper.pdf.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning, 2024. URL https://arxiv.org/abs/2310.12952.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

Hiroaki Sakoe. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:159–165, 1978. URL https://api.semanticscholar.org/CorpusID:17900407.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Lorenzo Steccanella and Anders Jonsson. State representation learning for goal-conditioned reinforcement learning. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 84–99. Springer, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Bailey Trang, Parham Saremi, Alan Q. Wang, Fangrui Huang, Zahra TehraniNasab, Amar Kumar, Tal Arbel, Li Fei-Fei, and Ehsan Adeli. Discovering latent graphs with GFlownets for diverse conditional image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PhHrlDKcx1.

Aleksandar Vakanski, Iraj Mantegh, Andrew Irish, and Farrokh Janabi-Sharifi. Trajectory learning for robot programming by demonstration using hidden markov model and dynamic time warping. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1039–1052, 2012. doi: 10.1109/TSMCB.2012.2185694.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*, pp. 36411–36430. PMLR, 2023.

Alan Wu, AJ Piergiovanni, and Michael S. Ryoo. Model-based behavioral cloning with future image similarity learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1062–1077. PMLR, 30 Oct–01 Nov 2020.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and KC Tan. Diversity-aware policy optimization for large language model reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=5eZ0iykpDU.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Training LLMs for divergent reasoning with minimal examples. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73115–73145. PMLR, 13–19 Jul 2025.

Xiaofeng Zhang, Aaron Courville, Michal Drozdzal, and Adriana Romero-Soriano. The intricate dance of prompt complexity, quality, diversity and consistency in t2i models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RBIBMCdw7y.

Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, kim hsieh, Tingjun Hou, and Jian Wu. Sample-efficient multi-objective molecular optimization with gflownets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79667–79684. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fbc9981dd6316378aee7fd5975250f21-Paper-Conference.pdf.

A Appendix

A.1 Calibration Of Hyperparameters.

Recall that the local similarity kernel used within GAK is $\kappa(s, s') = \exp(-d(s, s')/\sigma)$, where $\sigma > 0$ controls how aggressively the metric separates trajectories. A higher σ compresses similarities toward 1, making distinct trajectories appear more similar, while a lower σ amplifies differences. We select σ from the statistics of random rollouts.

As shown above, cosine normalization ensures $K_{ii} = 1$ and causes the combinatorial sum over alignment paths to approximately cancel between numerator and denominator (since these sums share the same set of valid paths for trajectories of similar length). The normalized similarity K_{ij} is therefore governed by a single representative alignment of length L. Denoting the median time-to-reach distance between two states from these rollouts by $\hat{d}$, we have that the expected distance for a median trajectory pair is $L \cdot \hat{d}$. We choose σ such that this median pair achieves a similarity of 0.5:

$$\exp\left(-\frac{L \cdot \hat{d}}{\sigma}\right) = 0.5 \quad \Longrightarrow \quad \sigma = \frac{L \cdot \hat{d}}{\ln 2}. \tag{7}$$

In practice, $\hat{d}$ is estimated by sampling 100 random rollouts of length $L = \text{median}(L_i)$, where L_i are the episode lengths of a baseline policy that solves the environment, computing all pairwise state distances, and taking their median. This calibration is performed once per environment. Alternatively, σ can be inherited from prior work on the same task. The Vendi score is then computed from this normalized kernel matrix, yielding the effective number of distinct trajectories in the population. Below are our calibrated σ values for the different environments.

Environment	L	$\hat{d}$ (median)	σ
Single-goal (small)	16	2.8	64.6
Single-goal (normal)	24	3.82	132.3
Single-goal (large)	32	4.75	219.3
Multi-goal Fig.3	12	3.5	60.6
Multi-key Fig.4	57	3	247
Single-goal (continuous) Fig.5	217	65	20349

The calibrated values reflect meaningful properties of each environment. Within the single-goal grid worlds, σ increases with environment size (64.6 $\rightarrow$ 132.3 $\rightarrow$ 219.3), capturing that larger environments admit longer trajectories and thus require a broader similarity bandwidth to distinguish genuinely different trajectories from minor deviations. Similarly, the multi-key environment yields a higher σ (247) than comparably sized single-goal settings, consistent with its longer episodes ($L = 57$) that arise from navigating to multiple subgoals, which in turn expand the space of possible trajectory variations. The continuous environment exhibits a substantially larger σ (20,349), reflecting both the finer-grained state space and the longer rollouts ($L = 217$), which together increase the range of pairwise distances and require correspondingly wider bandwidth. In all cases, the calibration adapts automatically through L and $\hat{d}$, requiring no manual tuning across qualitatively different environments.

A.2 Environment Variations

We verify that TVS remains interpretable as the environment structure varies. Fig. 7 shows single-goal environments of increasing size. The score remains consistent with the number of distinct strategies available, confirming that the automatic calibration of σ through L and $\hat{d}$ correctly adapts to environment scale.

Fig. 6 shows multi-goal environments where the number of available paths increases from 2 to 8. As each pair of new paths is introduced, TVS increases by approximately 2, closely tracking the number of genuinely distinct trajectories and converging to $\sim$8 when all eight paths are available.

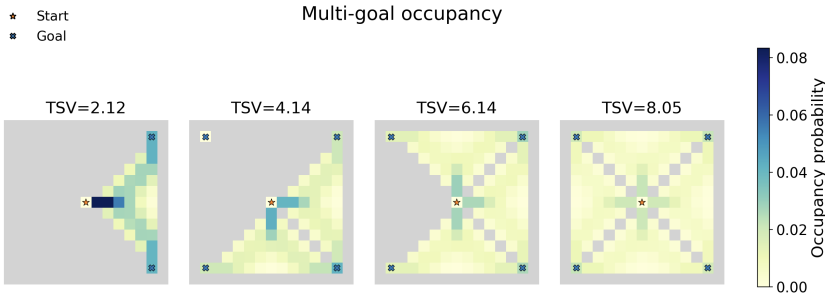

Figure 6: TVS across multi-goal environments with an increasing number of paths (from 2 to 8). The score increases by approximately 2 with each pair of new paths, reflecting the added behavioral diversity.

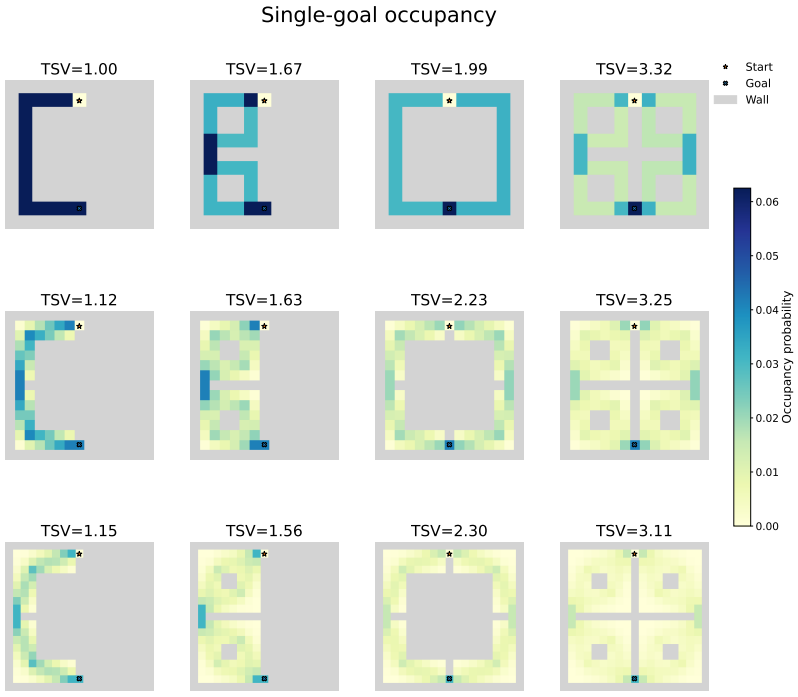

Figure 7: TVS across single-goal environments of increasing size. The score remains interpretable and consistent with the number of distinct strategies despite changes in maze dimensions.

A.3 ROBUSTNESS

As noted in Section 4.1, the time-to-reach distance does not guarantee that $\kappa/(1+\kappa)$ is positive semi-definite, though in practice the resulting negative eigenvalues are negligible (on the order of 10^{-5}). Since the Vendi score is defined via the entropy of the eigenspectrum, negative eigenvalues must be addressed before computation. Our default approach clamps them to zero. To verify that this choice does not introduce artifacts, we compare against an alternative correction: adding diagonal jitter $\mathbf{K}' = \mathbf{K} + \epsilon\mathbf{I}$ for $\epsilon \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, which shifts all eigenvalues upward and guarantees positive semi-definiteness. Figure 8 shows the relative change in TSV compared to the clamped baseline. For $\epsilon \leq 10^{-3}$, which far exceeds the magnitude of the observed negative eigenvalues, the score changes by at most 0.20% across all configurations. Only at $\epsilon = 10^{-1}$, several orders of magnitude larger than any eigenvalue violation, does the score shift appreciably ($\sim 21\%$). This confirms that clamping and jittering produce virtually identical results in the regime relevant to our kernel, and that the small PSD violations do not meaningfully affect the metric.

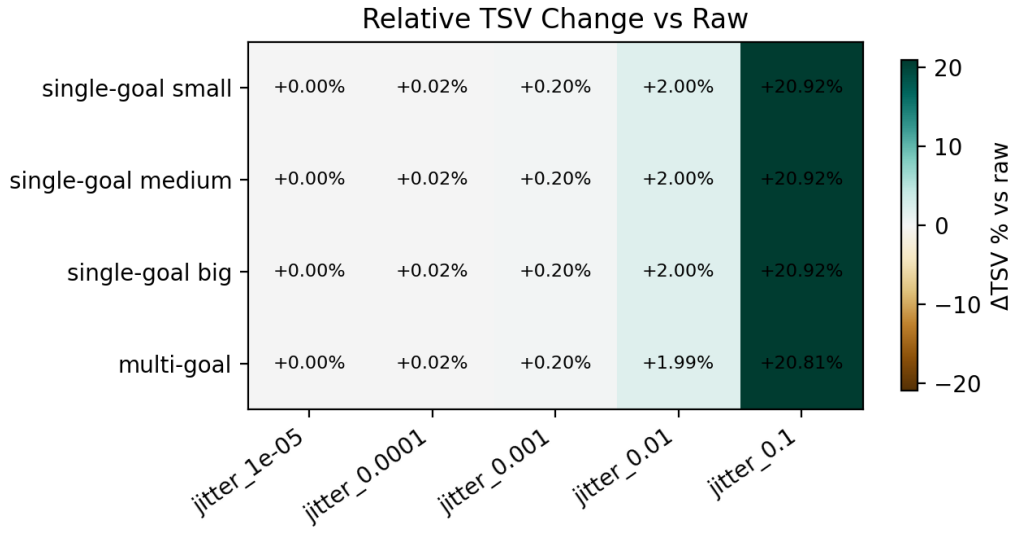

Figure 8: Relative change in Temporal Vendi Score when adding diagonal jitter $\epsilon\mathbf{I}$ to the kernel similarity matrix, compared to the raw (clamp) matrix.

In the continuous control setting, the negative eigenvalues are even smaller (on the order of 10^{-16}), consistent with numerical noise rather than structural PSD violations. The robustness pattern is similar: jittering with $\epsilon = 10^{-5}$ produces no measurable change, while $\epsilon = 10^{-4}, 10^{-3}, 10^{-2}$, and 10^{-1} yield relative shifts of approximately 0.02%, 0.2%, 2%, and 20% respectively, mirroring the discrete setting.

A.4 SCALING TO LARGER ENVIRONMENTS

We evaluate the impact of three computational optimizations on TVS: state discretization (rounding continuous observations to a fixed grid resolution), reducing the Sakoe–Chiba bandwidth below the default 20%, and subsampling trajectories by retaining every k-th state.

Fig. 9 reports the relative change in TVS compared to the baseline configuration (full precision, 20% band, no subsampling). State discretization to a resolution of 0.1 has negligible effect ($< 1\%$ relative change), validating its use in our continuous control experiments. Reducing the band to 15% also has minimal impact, but further reductions increasingly inflate the score as valid alignments are excluded. Subsampling has the opposite effect, systematically decreasing the score as the subsampling factor grows. Both effects are consistent across low and high-diversity agents.

Among these optimizations, state discretization is the most reliable for continuous control settings, as it preserves the full temporal structure while drastically reducing the number of unique distance evaluations. Reducing the Sakoe–Chiba bandwidth is likely environment-dependent: in mazes with simple structure, moderate reductions have little effect, but environments with more complex temporal dynamics may require wider bands. Subsampling is the most sensitive, as it alters the temporal resolution of trajectories. In particular, it may be problematic in environments with cyclic or repetitive behaviors, where the distinction between trajectories lies in short-term state ordering.

In terms of complexity, reducing the band from b to b' scales the per-pair GAK cost from $\mathcal{O}(T \cdot 2bT)$ to $\mathcal{O}(T \cdot 2b'T)$, a linear reduction. Subsampling by a factor k reduces both sequence lengths, yielding a quadratic speedup of $\mathcal{O}(T^2/k^2)$ per pair. To illustrate the practical impact of discretization on the one-time pairwise distance precomputation, evaluating 200 trajectories without discretization takes approximately one hour, whereas with discretization the full 1024-trajectory precomputation completes in under two minutes.

However, as shown in Fig. 5, the score stabilizes around 256 rollouts, for which computation takes under 5 minutes (see Fig. 10). More generally, convergence speed appears correlated with sample diversity: lower-diversity agents require fewer rollouts.

For settings where this cost remains prohibitive, the Vendi score supports Nyström approximation (Friedman & Dieng, 2023), which approximates the kernel matrix from a subset of columns. We leave this direction to future work.

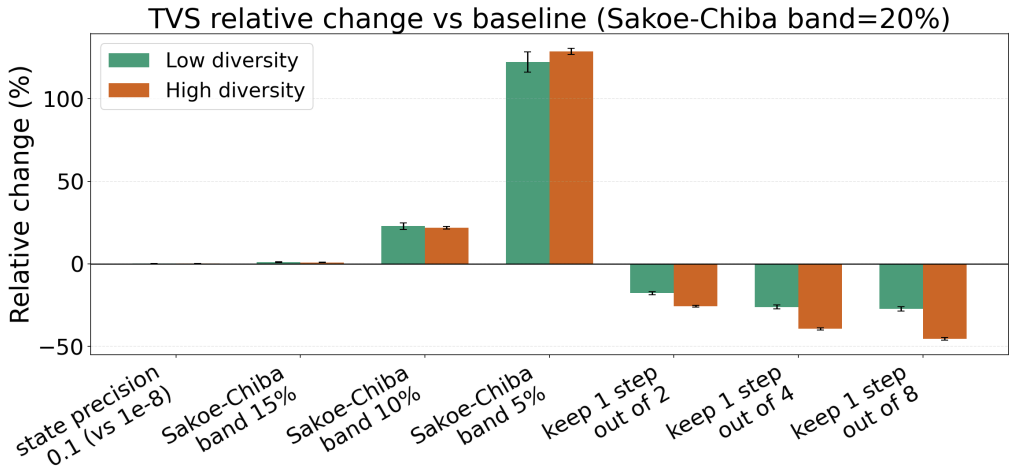

Figure 9: Relative change in TVS under three computational optimizations compared to the baseline configuration averaged over 4 seeds. (baseline : 200 trajectories, 10^{-8} state precision)

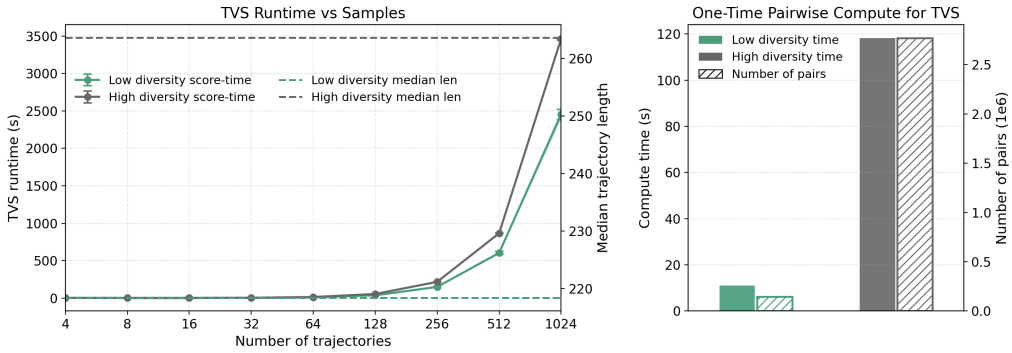

Figure 10: Computation cost of TVS in the continuous point maze with discretized states (resolution 0.1). **Left:** TVS runtime as a function of the number of sampled trajectories. Runtime scales quadratically in N due to pairwise trajectory comparisons. The high-diversity agent incurs higher cost owing to longer trajectories (median length $\sim$265 vs $\sim$218). **Right:** One-time precomputation cost. After discretization, the number of unique state pairs (hatched bars, right axis) and corresponding computation time (solid bars, left axis) depend on the agent's state-space coverage. Once precomputed, all subsequent GAK evaluations use table lookups.

A.5 ADDITIONAL EXPERIMENTS

Fig. 11 shows the pairwise distance matrix between trajectories (left) and the resulting normalized similarity matrix used by the Vendi score (right). Low distances correspond to high similarities, and the block structure visible in both matrices reflects the distinct behavioral clusters learned by the agent.

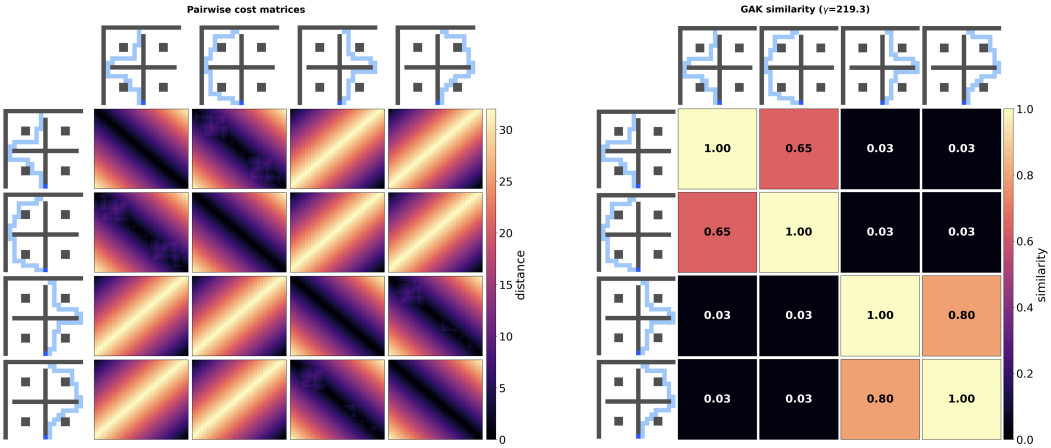

Figure 11: Pairwise trajectory distance matrix (left) and corresponding normalized similarity matrix (right).